Stat 602 Project Report

Kiegan Rice and Nate Garton

Due April 24, 2019

Introduction

The goal of this competition was to try to train an effective binary classifier on a small training set with many predictor variables. The criteria used to determine the best solution was the area under the ROC curve generated with predictions on the test set. The training data set had dimensions 250×300, and the test data had dimensions 19750×300.

Exploratory Data Analysis

Data Quality

The quality of the data was pristine. There were no missing values in the test or training data. There was some class imbalance in the training set (roughly 64% of the observations were class 1). A 95%, two-sided confidence interval based on large sample normality assumptions for the population class proportion was $(0.58, 0.70)$. Thus, assuming that we were truly given a random sample from the total data, it seemed unlikely that the classes were truly balanced. However, even at the upper bound of the confidence interval, the class imbalance would not likely be large enough to suggest problems with typical classification models/algorithms.

Class Conditional Distributions

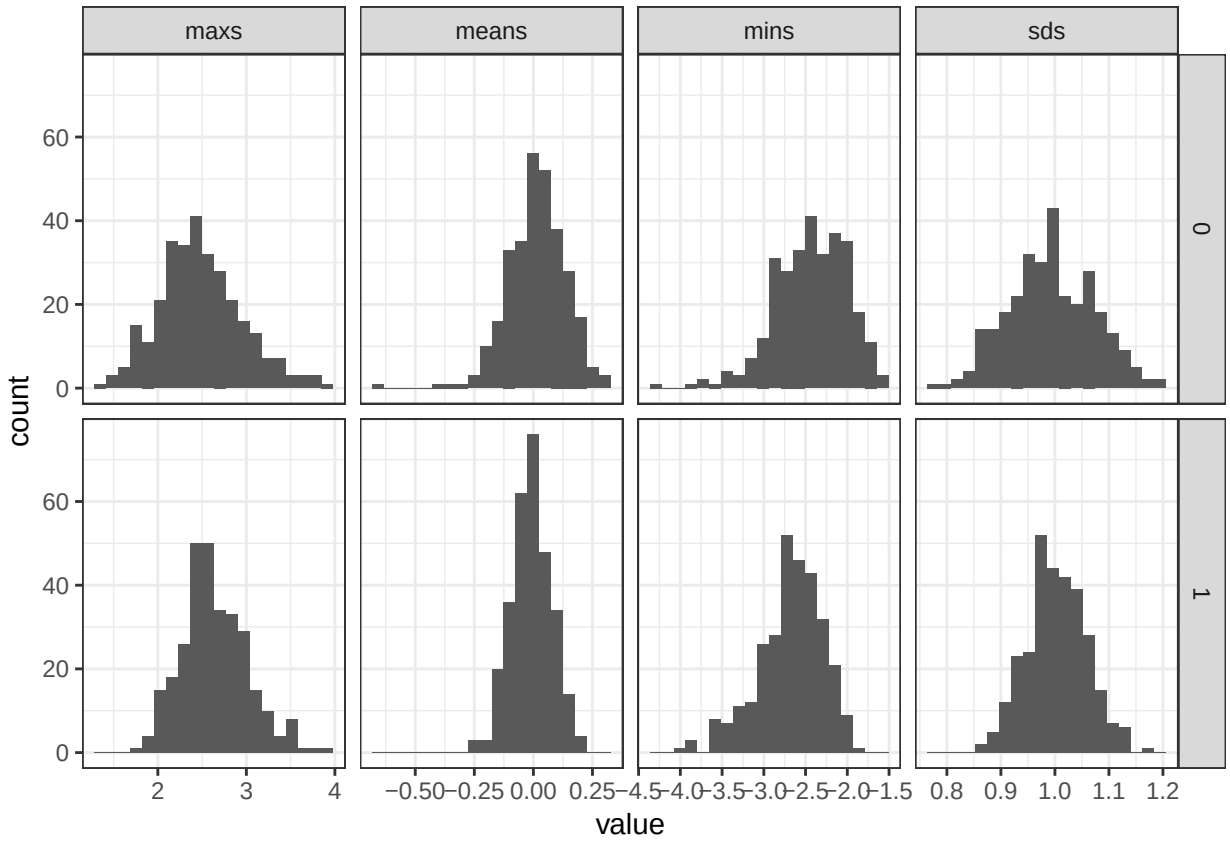

Figure 1: Class conditional histograms for column minimums, maximums, means, and standard deviations. Each row contains the column distribution for these four summary statistics for one of the two classes.